Metagenomic Analysis

An overview of the sequencing tools

The jargon

- Raw reads
- Unassembled vs assembled
- Read based
- K-mer
- Reference

The objecive

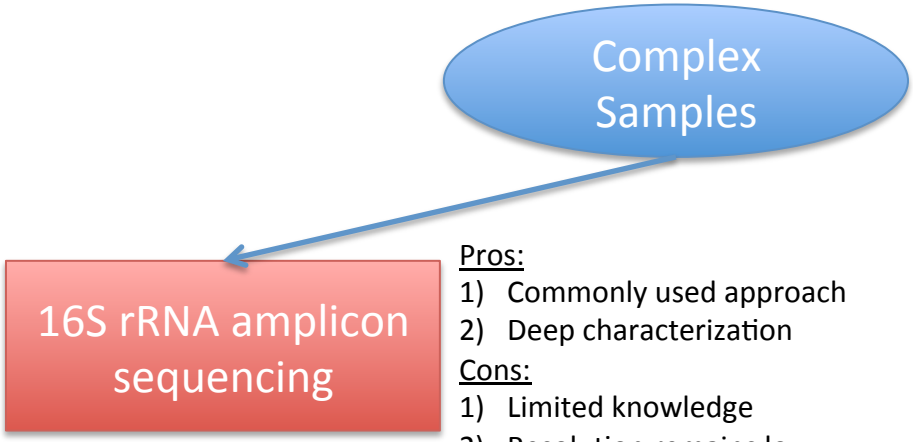
- 16S rRNA amplicon
  - OTU table (counts per “species”)
  - Classification (“species” taxonomy)
  - Metadata (sample specific information)

The objecive

- Metagenomic
  - Gene/genome count table (counts per “species”)
  - Classification (gene taxonomy/function)
  - Metadata (sample specific information)

Getting the most out of your data

ID, Abundance, Function



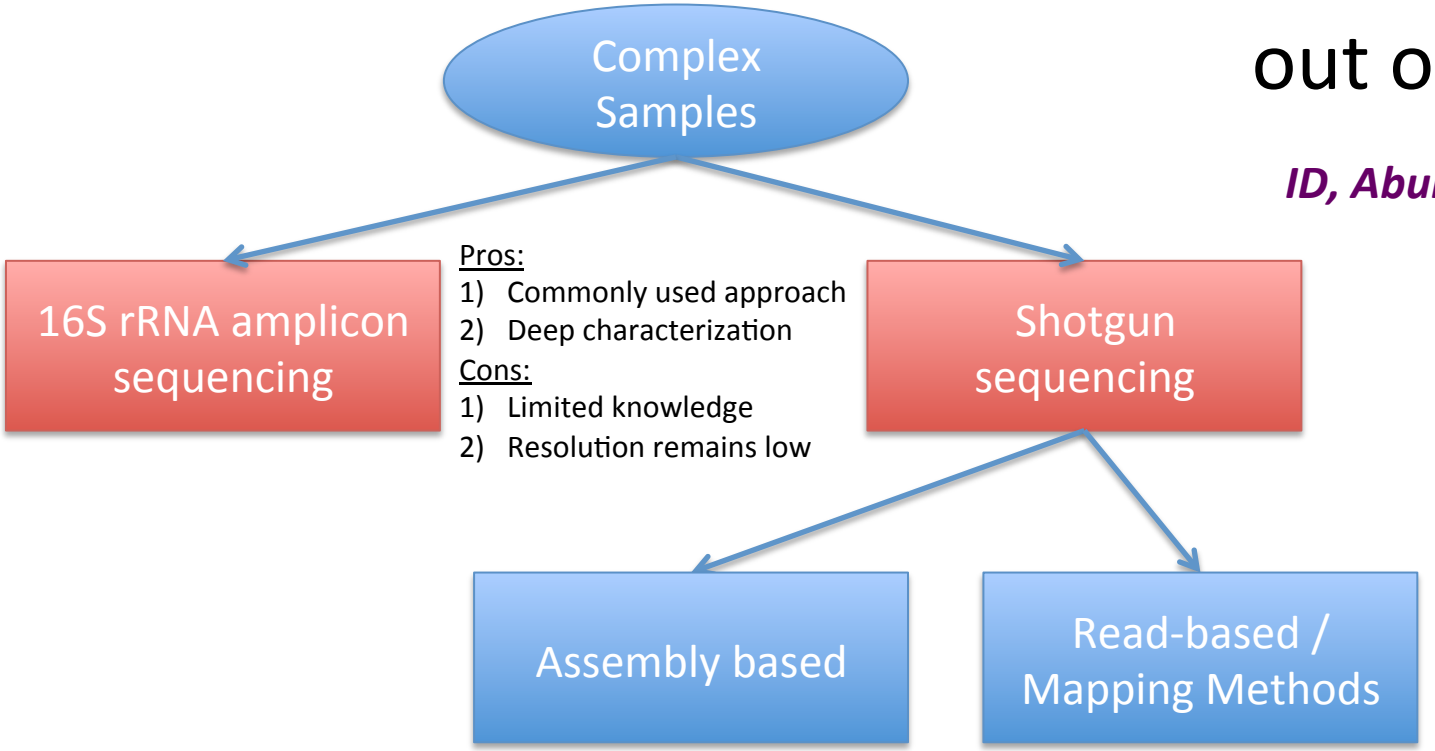
Pros:
1) Commonly used approach
2) Deep characterization

Cons:
1) Limited knowledge
2) Resolution remains low

Patrick Chain – props for these slides

Getting the most out of your data

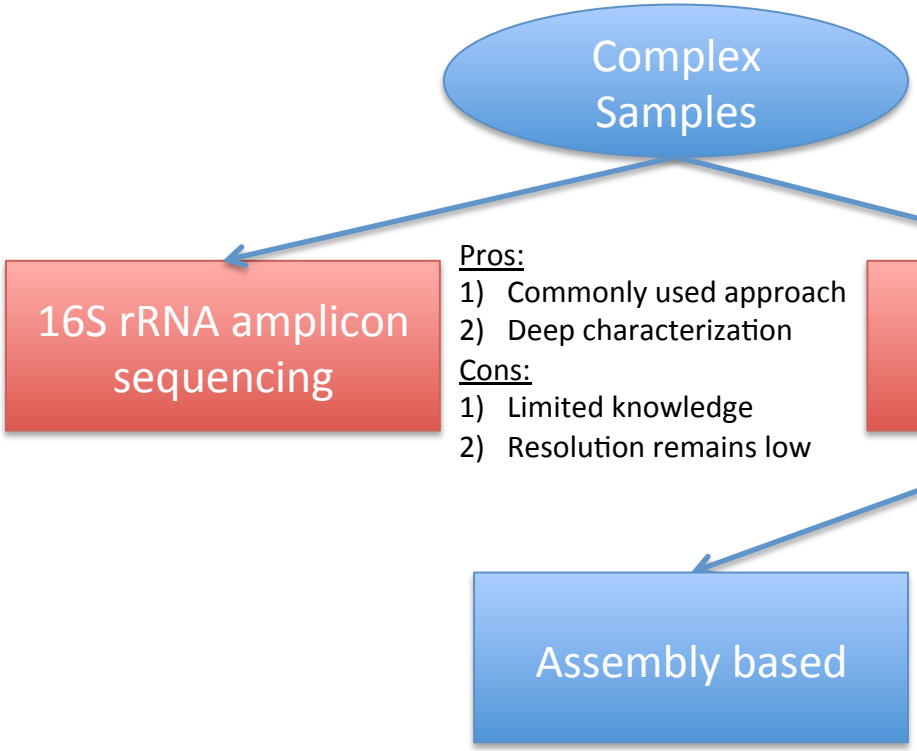
ID, Abundance, Function



Patrick Chain – props for these slides

Getting the most out of your data

ID, Abundance, Function



Pros:
1) Commonly used approach
2) Deep characterization

Cons:
1) Limited knowledge
2) Resolution remains low

Pros:
1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

Cons:
1) Massive data (short + with errors)
2) Lack of specificity due to FPs from genomic redundancy
3) Difficult to detect novel genomes – must infer

Patrick Chain – props for these slides

Getting the most out of your data

ID, Abundance, Function

16S rRNA amplicon sequencing

Pros:
1) Commonly used approach
2) Deep characterization

Cons:
1) Limited knowledge
2) Resolution remains low

Shotgun sequencing

Pros:
1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

Cons:
1) Massive data (short + with errors)
2) Lack of specificity due to FPs from genomic redundancy
3) Difficult to detect novel genomes – must infer

Assembly based

Pros:
1) Large contigs
2) Positional Information
3) Most direct method to identify novel orgs/genes

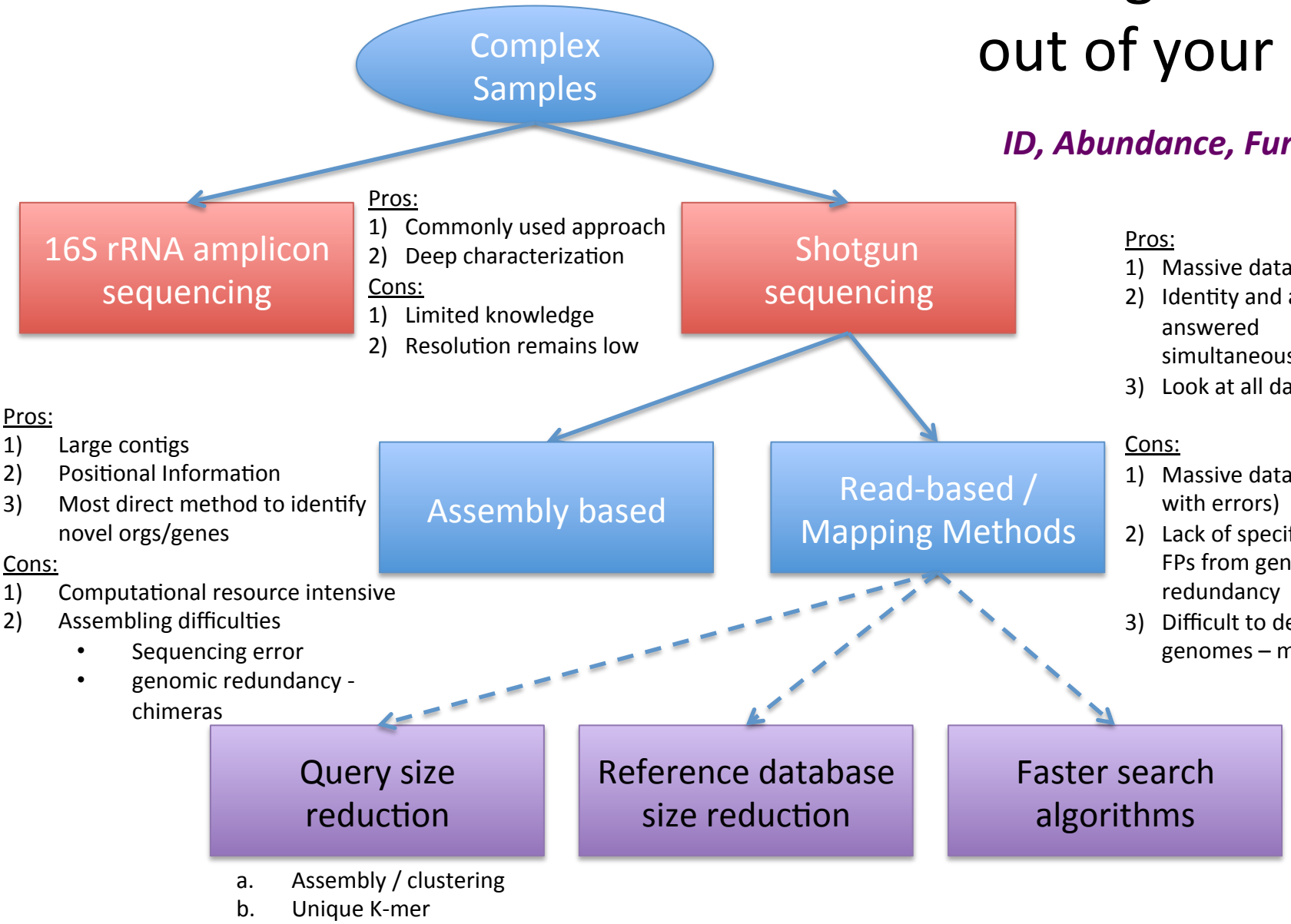
Cons:
1) Computational resource intensive
2) Assembling difficulties
   - Sequencing error
   - genomic redundancy - chimeras

Read-based / Mapping Methods

Patrick Chain – props for these slides

Getting the most out of your data

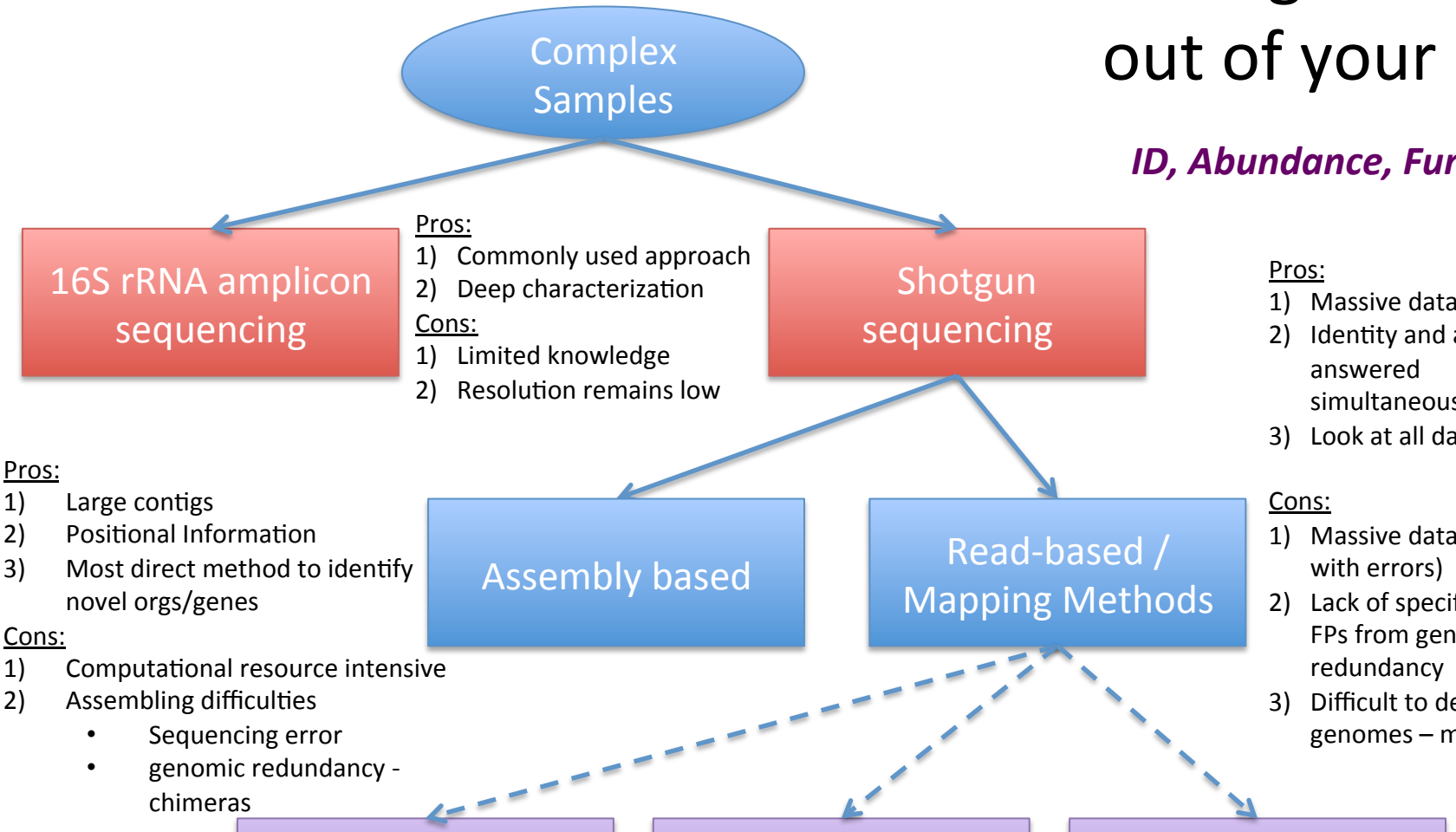
ID, Abundance, Function



Patrick Chain

Getting the most out of your data

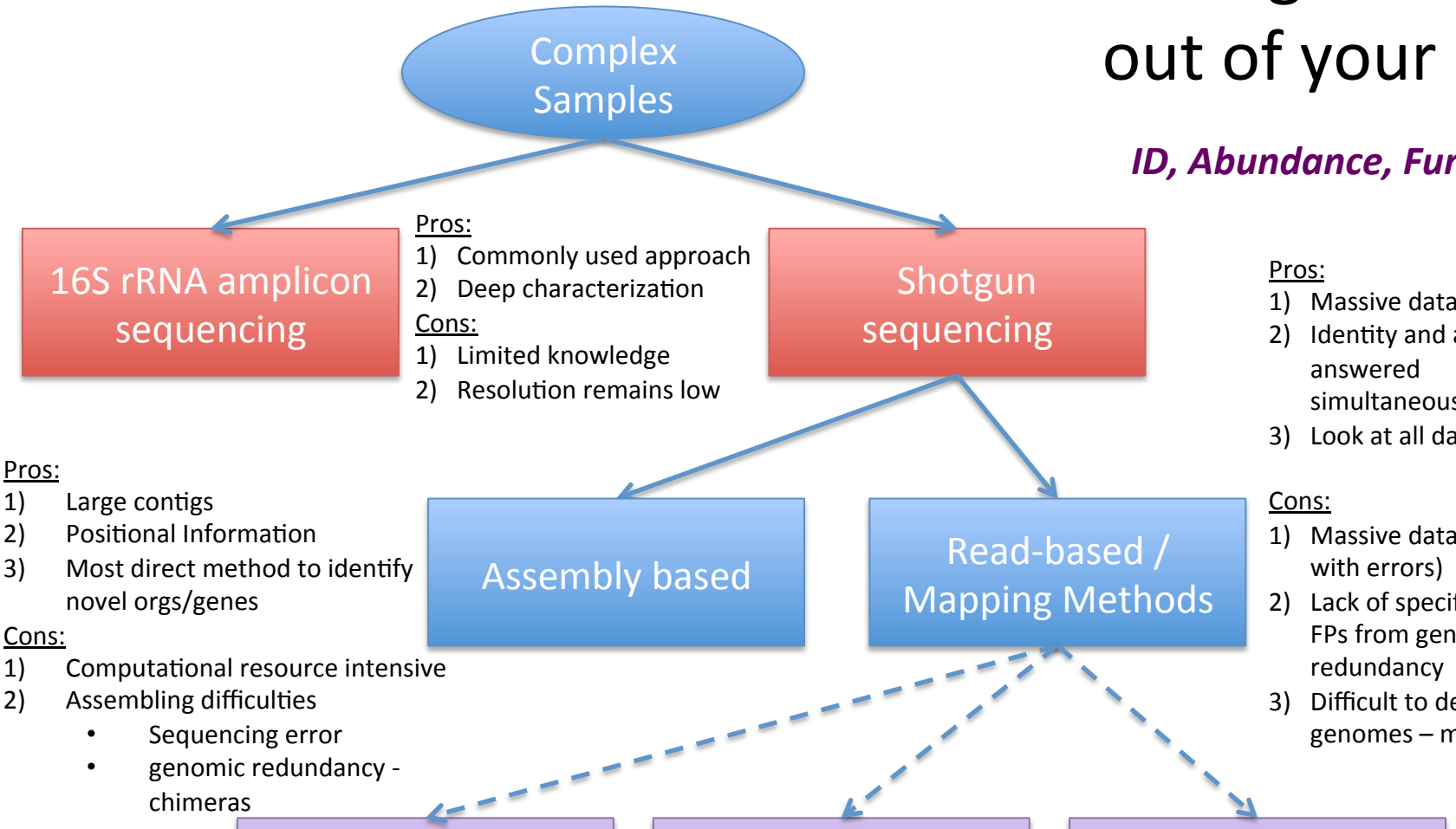
ID, Abundance, Function



Patrick Chain

Getting the most out of your data

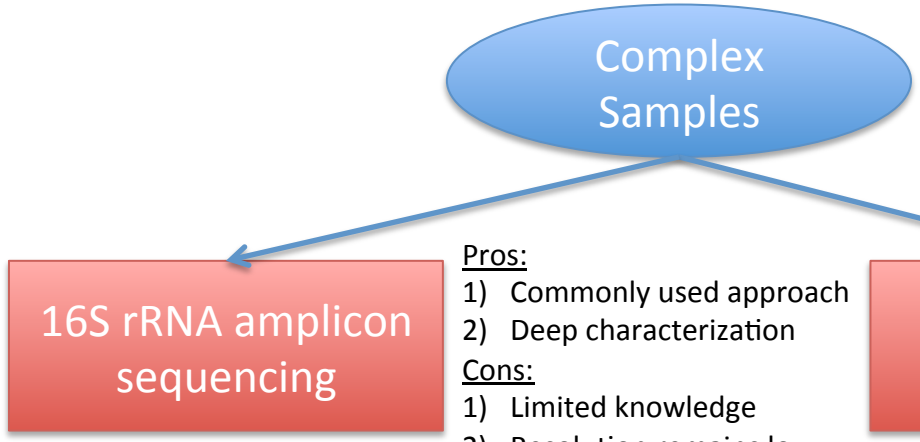
ID, Abundance, Function



Patrick Chain

Getting the most out of your data

ID, Abundance, Function



Pros:
1) Commonly used approach
2) Deep characterization

Cons:
1) Limited knowledge
2) Resolution remains low

Pros:
1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

Cons:
1) Massive data (short + with errors)
2) Lack of specificity due to FPs from genomic redundancy
3) Difficult to detect novel genomes – must infer

Pros:
1) Large contigs
2) Position
3) Most di
novel o

Cons:
1) Computational resource intensive
2) Assembling difficulties
   - Sequencing error
   - genomic redundancy - chimeras

Tutorial One: Quality trim your unassembled reads

Read-based / Mapping Methods

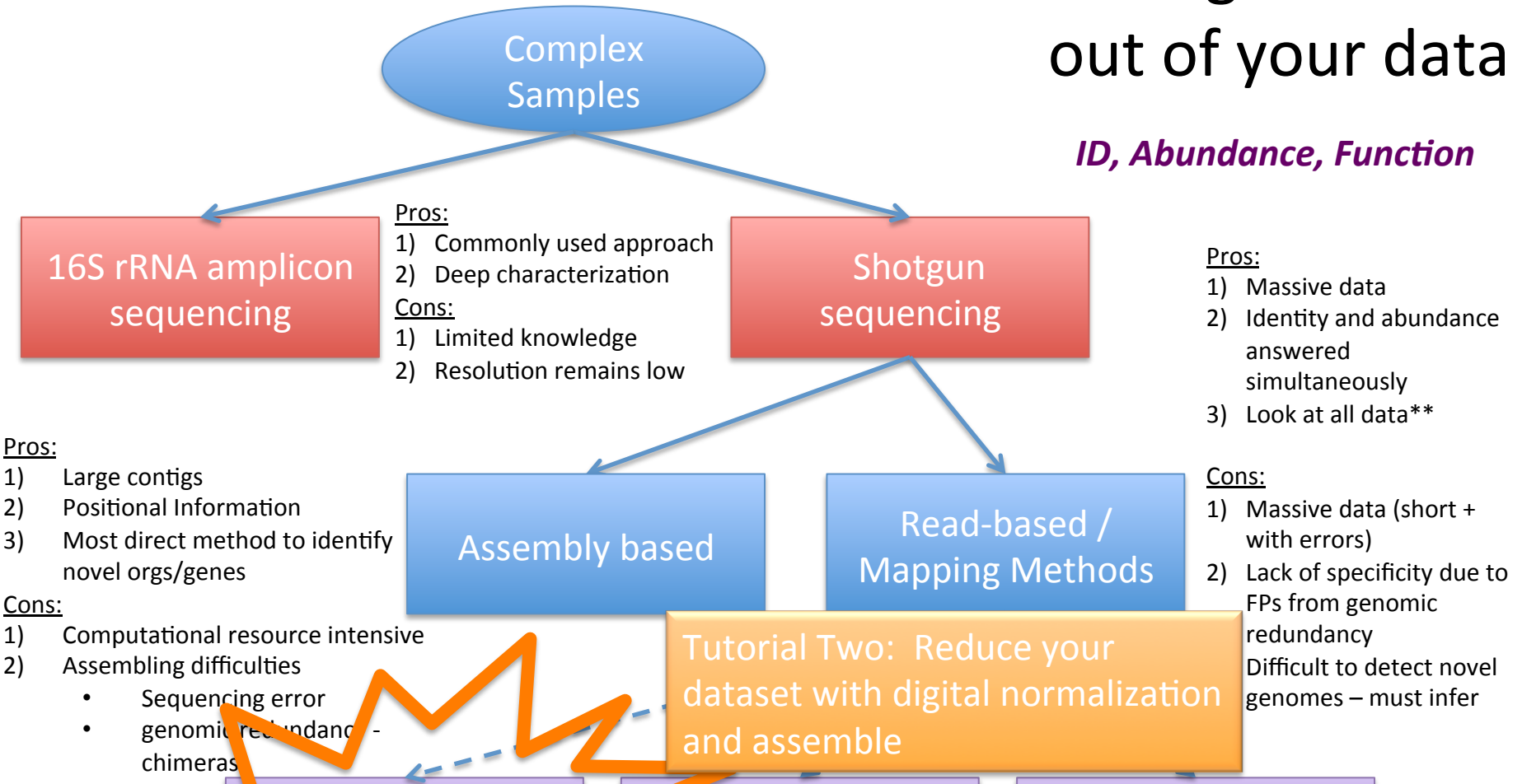
Query size reduction	Reference database size reduction	Faster search algorithms

a. Assembly / clustering
b. Unique K-mer
	a. Selection of marker genes
	b. Identification of signatures (Kmers)
		a. Exact match
		b. K-mer based search
		c. Improved algorithm
		  a. Clustering
		  b. Pattern matching

Patrick Chain

Getting the most out of your data

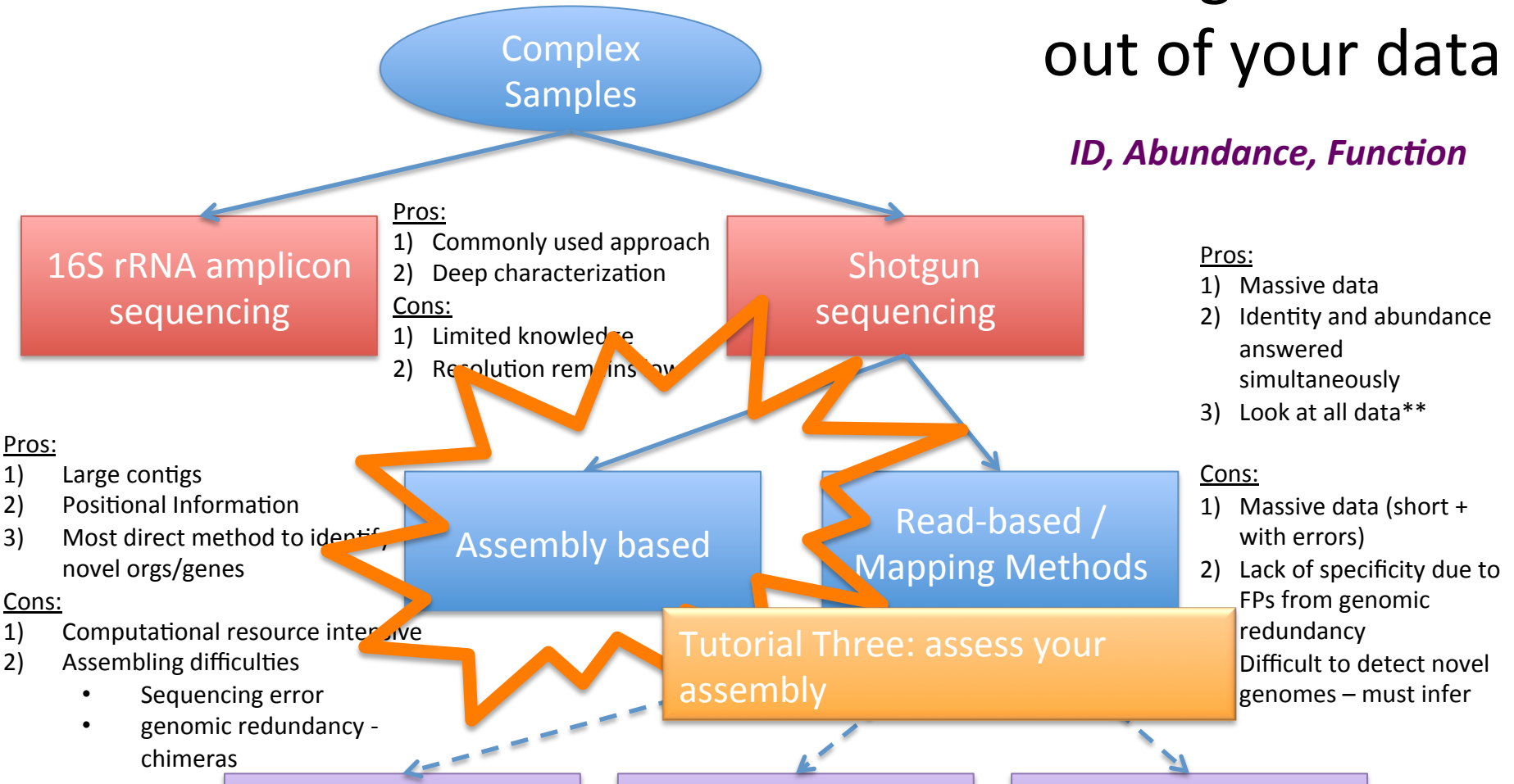
ID, Abundance, Function



Patrick Chain

Getting the most out of your data

ID, Abundance, Function



Pros:
1) Commonly used approach
2) Deep characterization

Cons:
1) Limited knowledge
2) Resolution remains low

Pros:
1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

Cons:
1) Massive data (short + with errors)
2) Lack of specificity due to FPs from genomic redundancy
Difficult to detect novel genomes – must infer

Pros:
1) Large contigs
2) Positional Information
3) Most direct method to identify novel orgs/genes

Cons:
1) Computational resource intensive
2) Assembling difficulties
   - Sequencing error
   - genomic redundancy - chimeras

a. Assembly / clustering
b. Unique K-mer

a. Selection of marker genes
b. Identification of signatures (Kmers)

a. Exact match
b. K-mer based search
c. Improved algorithm
   a. Clustering
   b. Pattern matching

Patrick Chain

Getting the most out of your data

ID, Abundance, Function

16S rRNA amplicon sequencing

Pros:
1) Commonly used approach
2) Deep characterization

Cons:
1) Limited knowledge
2) Resolution remains low

Shotgun sequencing

Pros:
1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

Cons:
1) Massive data (short + with errors)
2) Lack of specificity due to FPs from genomic redundancy
3) Difficult to detect novel genomes – must infer

Assembly based

Pros:
1) Large contigs
2) Positional Information
3) Most direct method to identify novel orgs/genes

Cons:
1) Computational resource intensive
2) Assembling difficulties
   - Sequencing error
   - genomic redundancy - chimeras

Read-based / Mapping Methods

Tutorial Four: Mapping reads to estimate abundances

Query size reduction

a. Assembly / clustering
b. Unique K-mer

Reference database size reduction

a. Selection of marker genes
b. Identification of signatures (Kmers)

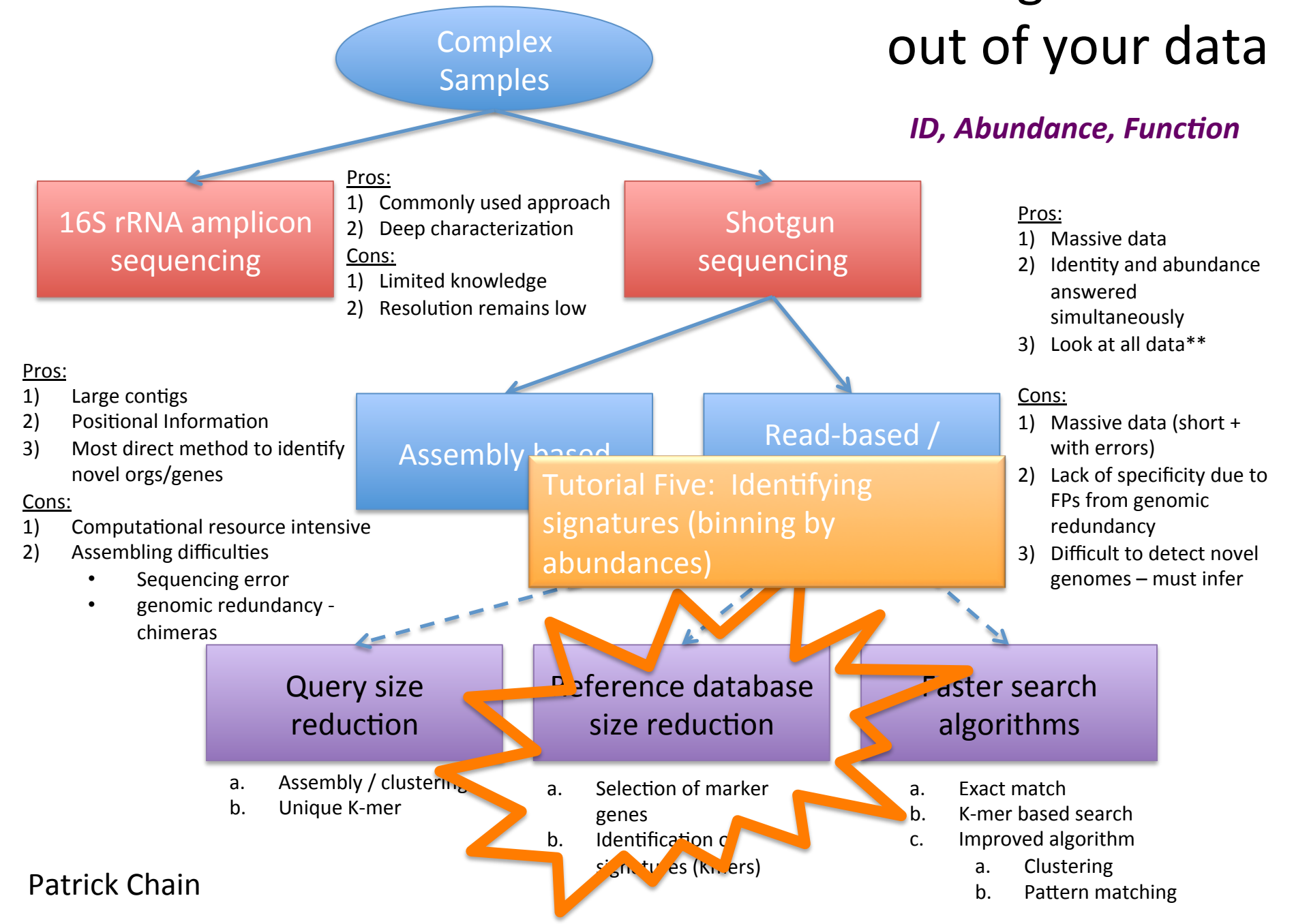
Faster search algorithms

a. Exact match
b. K-mer based search
c. Improved algorithm
   a. Clustering
   b. Pattern matching

Patrick Chain

Getting the most out of your data

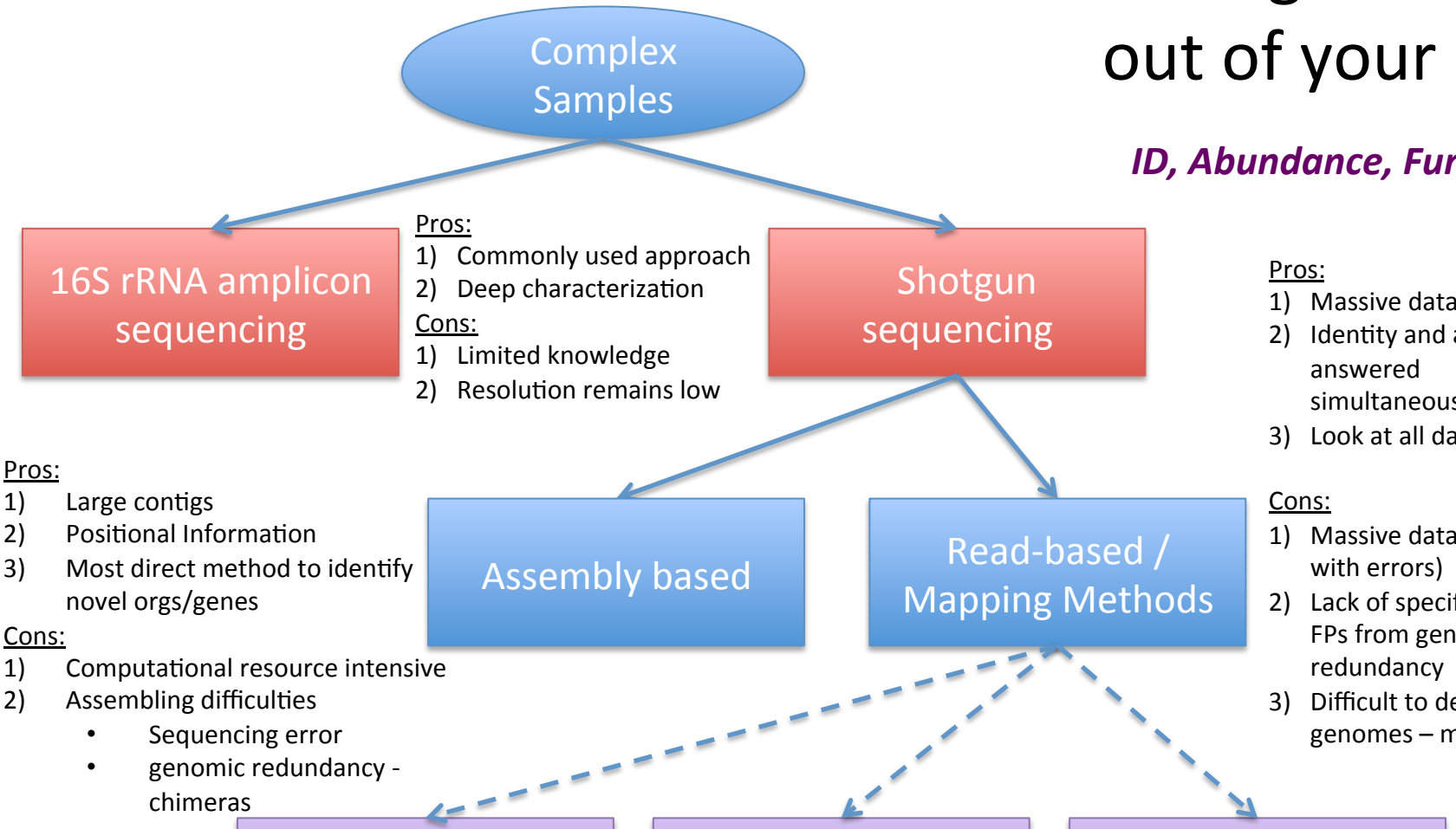
ID, Abundance, Function



Patrick Chain

Getting the most out of your data

ID, Abundance, Function



Patrick Chain

So let’s do some learning…